Report: Comparison of Gradient Descent Methods for Linear Regression

1. Introduction

In this assignment, we implemented and compared three variants of gradient descent—**Stochastic Gradient Descent (SGD)**, **Batch Gradient Descent**, and **Mini-Batch Gradient Descent**—for training a simple linear regression model. We also experimented with different learning rates and momentum values to observe their impact on convergence. This report summarizes our observations, challenges faced, and conclusions drawn from the experiments.

2. Observations

2.1 Performance of SGD, Batch, and Mini-Batch Gradient Descent

Stochastic Gradient Descent (SGD):

- **Convergence Speed**: SGD converges quickly but with high variance in the loss. The updates are noisy because each update is based on a single data point.
- **Final Parameters**: The final parameters are close to the true values but may fluctuate due to the noisy updates.
- **Use Case**: Suitable for large datasets where computing the full gradient is expensive.

Batch Gradient Descent:

- **Convergence Speed**: Batch GD converges slowly but smoothly because it uses the entire dataset for each update.
- **Final Parameters**: The final parameters are stable and accurate.
- **Use Case**: Suitable for small datasets where computing the full gradient is feasible.

Mini-Batch Gradient Descent:

- **Convergence Speed**: Mini-Batch GD strikes a balance between SGD and Batch GD. It converges faster than Batch GD and is less noisy than SGD.
- **Final Parameters**: The final parameters are stable and close to the true values.
- **Use Case**: Suitable for medium to large datasets where a balance between speed and stability is required.

2.2 Impact of Learning Rates and Momentum

Learning Rates:

- **Small Learning Rate (e.g., 0.001)**:
  - Convergence is slow but stable.
  - The model may get stuck in local minima if the learning rate is too small.
- **Large Learning Rate (e.g., 0.1)**:
  - Convergence is faster but may overshoot the optimal solution.
  - The loss may diverge if the learning rate is too large.
- **Optimal Learning Rate (e.g., 0.01)**:
  - Balances speed and stability, leading to smooth and fast convergence.

Momentum:

- Momentum accelerates convergence by adding a fraction of the previous update to the current update.
- It helps reduce oscillations in the loss, especially in SGD.
- A momentum term of 0.9 worked well in our experiments, leading to faster and smoother convergence.

3. Challenges

1. **Hyperparameter Tuning**:
   - Choosing the right learning rate and momentum term was challenging. Too small a learning rate led to slow convergence, while too large a learning rate caused divergence.
   - For Mini-Batch GD, selecting an appropriate batch size required experimentation.
2. **Noisy Updates in SGD**:
   - The noisy updates in SGD made it difficult to achieve stable convergence. Adding momentum helped mitigate this issue.
3. **Implementation Details**:
   - Ensuring correct gradient calculations and parameter updates required careful debugging.
   - Handling edge cases, such as division by zero or NaN values, was necessary.

4. Conclusions

1. **Gradient Descent Methods**:
   - SGD is fast but noisy, making it suitable for large datasets.
   - Batch GD is slow but stable, making it ideal for small datasets.
   - Mini-Batch GD provides a good trade-off between speed and stability, making it a popular choice for most applications.
2. **Learning Rates**:
   - The learning rate is a critical hyperparameter that significantly impacts convergence. An optimal learning rate balances speed and stability.
3. **Momentum**:
   - Momentum is a powerful technique to accelerate convergence and reduce oscillations, especially in SGD.
4. **Final Model Parameters**:
   - All three methods (SGD, Batch GD, Mini-Batch GD) converged to similar final parameters, demonstrating the correctness of the implementation.

5. Figures and Plots

Below are the key plots generated during the experiments:

1. **Loss vs Epochs for SGD, Batch GD, and Mini-Batch GD**:
   - This plot compares the convergence behavior of the three methods.
2. **Impact of Learning Rates**:
   - This plot shows how different learning rates affect convergence.
3. **Impact of Momentum**:
   - This plot demonstrates the effect of momentum on SGD convergence.